When the Crowd Challenges the Lab: Lessons Learnt from Subjective Studies on Image Aesthetic Appeal

Judith Redi
TU Delft
j.a.redi@tudelft.nl

Ernestasia Siahaan
TU Delft
E.Siahaan@tudelft.nl

Pavel Korshunov
Ecole Polytechnique Fédérale de Lausanne
pavel.korshunov@epfl.ch

Julian Habigt
Technische Universität München
jh@tum.de

Tobias Hossfeld
University of Duisburg-Essen
tobias.hossfeld@uni-due.de

ABSTRACT

Crowdsourcing gives researchers the opportunity to collect subjective data quickly, in the real-world, and from a very diverse pool of users. In a long-term study on image aesthetic appeal, we challenged the crowdsourced assessments with typical lab methodologies in order to identify and analyze the impact of crowdsourcing environment on the reliability of subjective data. We identified and conducted three types of crowdsourcing experiments that helped us perform an in-depth analysis of factors influencing reliability and reproducibility of results in uncontrolled crowdsourcing environments. We provide a generalized summary of lessons learnt for future research studies which will try to port lab-based evaluation methodologies into crowdsourcing, so that they can avoid the typical pitfalls in design and analysis of crowdsourcing experiments.

1. INTRODUCTION

Crowdsourcing (shortened as CS in the rest of the paper) may be a game changer for Quality of Experience (QoE) research. Aimed at measuring (unobtrusively and automatically) users' delight or annoyance with a multimedia application or service [15], research on QoE heavily relies on the deployment of user studies, which are crucial for understanding the mechanisms underlying user appreciation for multimedia experiences [20]. Traditionally, these studies are performed in a highly controlled lab setting to ensure the reliability of the collected data and the repeatability of the results. However, this rigid environment poses time and cost constraints, leading eventually into a limitation in the number of subjects and stimuli (i.e., images, videos, etc.) that can be involved in the user assessments [6]. CS, on the other hand, has the potential to reach a larger and more diverse group of subjects in a short timespan and at a low economical cost [14]. For this reason, the whole multimedia community is regarding CS with growing interest to be used for ground truth annotation (e.g., as per [10]). Similarly, QoE researchers are considering it as an alternative platform to carry out user studies on quality appreciation efficiently [9].

The enthusiasm around crowdsourcing has led to a plethora of QoE studies related to it, including [1, 13]. But along with success stories [9], CS has started showing its limitations [18]. The lack of control on environmental conditions, on subject commitment and understanding of the experimental task, significantly influences the reliability of the QoE measurements [7]. More importantly, the tendency to adapt typical lab-based experiments to CS without taking into account its intrinsic characteristics can lead to undesired outcomes.

In this paper, we identify and conduct three crowdsourcing experiments, representing typical attempts of replicating a lab-based study. An assessment of aesthetic appeal and recognizability of images was selected as an example of such study [19]. Through a thorough analysis of the CS 'replicas', including their design peculiarities and their resulting assessments, we aim to identify the constraints of the CS testing environment, those potential mistakes that unexperienced researchers can make when designing such experiments, and good approaches to ensure data reliability.

In particular, we investigate a set of hypotheses on what can potentially bring about discrepancies and irregularities in the outcome of crowdsourcing-based test compared to established and controlled lab-based experiment. Firstly, we explore two issues specifically related to CS, namely:

- *Bias due to the scoring task.* The impact of assessing multiple quantities (in our case, recognizability and aesthetic appeal) at a time.
- *Contextual effects.* The effect of contextual changes on the Mean Opinion Scores [3, 4], for instance, due to the fact that crowdsourcing-based assessment is usually fragmented into many small tasks, as opposed to the lab-based assessment.

Secondly, we investigate issues related to lab-to-CS repeatability, namely:

- *Instructions and scoring scale layout.* The importance of using exactly the same scoring scale and instructions in CS and lab to obtain repeatable results.

CrowdMM'15, October 30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3746-5/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/XXXXXXX.XXXXXXX.

- *Non-repeatable measurements.* The possibility that some measurements are not easily reproducible in the crowdsrourcing (or in any) environment.

2. RELATED WORK

Research on QoE relies on subjective tests, in which users are asked to sort stimuli (i.e., media) according to their perceived properties or attributes [6] on a given scale. Traditionally these tests have been performed in highly controlled, standardized environments [11] (typically within laboratory facilities), to allow minimizing the effect of:

- environmental factors, by controlling the lighting and viewing position, thereby making the visibility conditions homogeneous across participants
- misunderstandings on the experimental task, by having a test supervisor giving detailed instructions during a training phase and providing supplementary explanations on the task in case of need.

Along with high controllability, lab-based environment poses constraints in terms of the number of subjects and stimuli that can be used in the evaluations, since subjects are usually highly paid and can only spend limited time (between thirty minutes and one hour, to avoid fatigue effects [6,11]) on the test. In addition, the direct recruitment of participants in the vicinity of the lab location typically leads to low diversity in user demographics, which should instead be privileged [15].

Crowdsourcing [5] seems to be a promising solution to overcome the limitations of lab testing. For QoE assessments, subjective tests can be completed relatively quickly using CS, sometimes within a few minutes. The large crowd allows to easily investigate various test conditions and includes the real-life environment into the assessment, and with rather different demographics [8].

Besides the promising advantages of performing subjective tests in Crowdsourcing, there are still challenges that need to be addressed. The Internet-based, remote conduction of subjective tests is limited by technical factors such as bandwidth constraints, or support of the workers' software and devices to present the required stimuli. Moreover, as it is impossible to properly supervise the workers in the experimental tasks, the results are prone to errors in task understanding and sloppiness. Finally, CS tasks should be fairly short (up to 10 minutes) to avoid boredom and unreliable behavior. Thus, to collect QoE assessments for a large set of stimuli, experimenters usually have to decompose the test into a set of smaller tasks (i.e., campaigns), each one including a sub-set of the stimuli, increasing the risk of context effects [3, 4]. To tackle these challenges, suggestions related to payment scheme, worker selection, and task design have been extensively discussed, for example in [9].

Still, the question remains open as to what extent crowdtesting impacts QoE assessements and whether it can replace lab-based testing at all. In fact, several studies have compared QoE assessments obtained in lab and CS, and highlighted how lab-based assessments are still more consistent across users and reliable than those obtained through CS. This was shown in [2] for assessments of QoE for videos with different codecs and compression levels, as well as loss concealment schemes for IPTV. In [9], a similar conclusion was drawn for assessments of visual quality of H.264/AVC videos in lab compared with in CS. In [18], it was also shown that there were discrepancies when it comes to correlations of image recognizability and image aesthetic appeal in lab and CS experiments.

3. AN EXAMPLE STUDY: AESTHETIC APPEAL AND RECOGNIZABILITY

Understanding users appreciation for image aesthetic appeal is of major interest for image QoE assessment [20], as well as for several multimedia applications, from retrieval to automatic photo-editing [12]. Often considered too much depending on user factors ("beauty is in the eye of the beholder"), image aesthetic appeal has been shown to be quantifiable within confidence intervals just slightly larger than those usually obtained in lab-based investigations of image perceptual quality [20]. It was therefore interesting to verify whether this was also true when performing the assessments in a less controlled crowdsourcing setting. We set out to repeat in a CS environment an existing lab experiment [19], investigating (1) aesthetic appeal and (2) recognizability of the content of images. This second quantity is related to perceptual fluency [16], known to have an effect on the aesthetic appeal of works of art. In this study, we wanted to check whether this effect was preserved when judging the aesthetic appeal of consumer images.

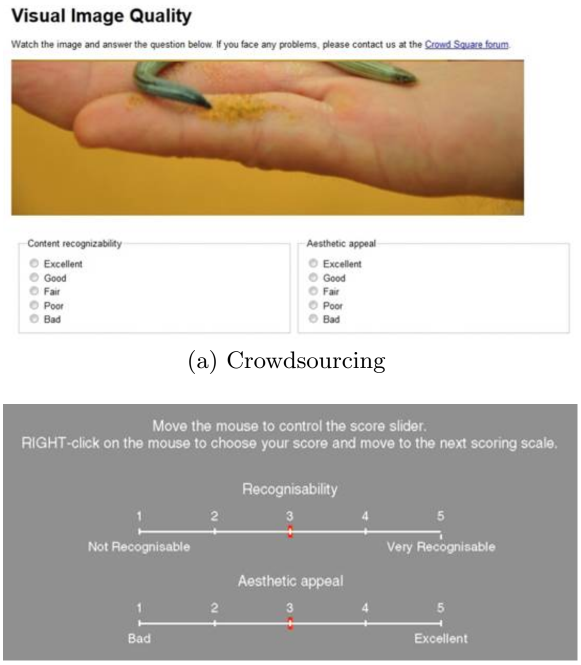
3.1 Experimental methodology

Lab experiment. Nineteen subjects, all studying or working in university, assessed the aesthetic appeal and recognizability of 200 consumer images spanning different content categories. Subjects were initially briefed about the general setup and their task, and went through a training session consisting of the scoring of three images with different levels of aesthetic appeal and recognizability. The training helped ensure (1) that the participant understood the two different tasks, and (2) that for both tasks s/he had formed some visual reference to function as an anchor for the usage of the scales [6]. During the test, the aesthetic appeal assessment of images were performed in a Single Stimulus setup. Particpants had to score aesthetic appeal on a 5-point ACR scale, and recognizability on a 5-point discrete scale ranging from "Not Recognizable" to "Very recognizable" (see Fig. 1b). Scales were shown on a separate screen from the image to avoid distraction. Due to the large number of images, the experiment was split into 4 sessions, including 50 images each, between which participants could take a short break. The experimental set-up followed the ITU-R BT.500 recommendation [11] and no time constraint was given for image observation and scoring.

CS experiment. To transpose this experiment into a crowdsourcing setting, we had to make several amends to the experimental protocol.

1. Images were assessed in 13 campaigns of 20 images each. This was necessary to make sure that a task duration would last less than five minutes, which is recommended for QoE experiments in CS [9].
2. Five of the 20 images were kept equal for all campaigns. They were chosen to have aesthetic appeal scores corresponding to the 0th, 25th, 50th, 75th, and 100th per-

[0] A sample of the images can be found in the supplemental materials submitted with this paper, and a description of the criteria used for image selection is specified in [19]



Visual Image Quality

Watch the image and answer the question below. If you face any problems, please contact us at the Crowd Square forum.

Content recognizability

○ Excellent
○ Good
○ Fair
○ Poor
○ Bad

Aesthetic appeal

○ Excellent
○ Good
○ Fair
○ Poor
○ Bad

(a) Crowdsourcing

Move the mouse to control the score slider.
RIGHT-click on the mouse to choose your score and move to the next scoring scale.

(b) Laboratory

Figure 1: Scoring interfaces as used in the CS and lab experiments.

centiles of the distribution of all aesthetic quality MOS, as observed in the lab assessment of the 200 images. We will refer to these images as "anchors". They had the purpose of limiting contextual effects [3] by fixing the extreme values of aesthetic appeal to be seen in each campaign.

3. To allow filtering unreliable participants, we introduced control questions at different points of the test. The questions asked users the content they saw in the previous image in a multiple choice form.
4. We used the QualityCrowd [13] framework due to its flexibility and therefore easy adaptation to the task of aesthetics and recognizability assessment. Ergo, the scoring interface could not be kept identical to that used in the lab experiment. In addition, the recognizability scale had to be converted into an ACR scale (see Fig. 1a).
5. Due to implementation constraints, both questions had to be displayed on the same page as the image being assessed; specifically, the recognizability question was displayed on the left and the aesthetic appeal one on the right, both below the image (see Fig. 1a).

The training session was maintained, similarly to the lab test, at the beginning of the session. For each of the 13 campaigns, 30 participants (workers) assessed 20 images for 0.30 USD. Data were collected within one week. Before proceeding with the analysis of the scores, unreliable workers were filtered out by eliminating workers answering the content questions incorrectly, taking an unusual amount of time to score the images, or scoring randomly (according to the outlier detection in [11]). Although we reached out to workers from diverse geographical locations, in this paper we refer only to data relative to north-american workers, who were shown to answer the content questions correctly most often [18].

3.2 Results

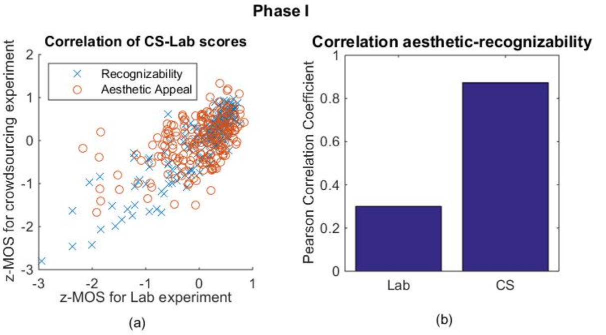


Figure 2: Results of Phase I: (a) correlation between z-MOS obtained in Lab and CS for Recognizability and Aesthetic Appeal. (b) Correlation between z-MOS of aesthetic appeal and recognizability in Lab and CS.

To check the consistency between aesthetic appeal and recognizability assessments collected in the Lab and CS, we examined the correlation of the respective Mean Opinion Scores (i.e. the average of the scores given by all participants to the same image). Whereas recognizability MOS were found to be strongly correlated between Lab and CS ($\rho = 0.87$), this was not the case for aesthetic appeal ($\rho = 0.51$), as shown in Fig. 2.a. Additionally, we checked whether the relationship between aesthetic appeal and recognizability was preserved from Lab to CS. We computed the correlation between MOS of aesthetic appeal and MOS of recognizability, separately for Lab and CS. Results are shown in Fig. 2.b. Whereas in the CS setting recognizability MOS were predictive for aesthetics MOS, this was not the case for the Lab data. Though we verified repeatablity of Lab-based Recognizability measurements in CS, we could not confirm the same for Aesthetic Appeal assessments.

3.3 Discussion

We identified four possible causes for the lack of repeatability of aesthetic appeal measurements:

Bias due to the scoring interface. Participants scored aesthetic appeal and recognizability simultaneously. Priming effects, poor instructions (e.g., the terminology "aesthetic appeal" was unclear), or simply carelessness, could have led CS workers to interpret the two attributes as related and scored them similarly. The layout of the CS scoring interface (see Fig.1), where the left question (recognizability) was naturally answered first may also induce this interpretation. Workers possibly scored recognizability truthfully, and replicated the same judgment on the aesthetic appeal scale (right), for ease of completion (which is not uncommon in CS [9]) or due to unconscious bias. We hypothesize therefore that, *by asking workers to rate only aesthetic appeal rather than both attributes*, we could obtain aesthetic appeal MOS closer to those obtained in the lab and less similar to the recognizability MOS.

Contextual effects. Despite the use of anchor images, context effects [3] may still have affected the MOS in CS, due to the division of the image set in 13 campaigns. We hypothesize therefore that *by re-aligning the MOS across all campaigns, by means of the anchors values*, we could eliminate context effects and thereby the discrepancies between Lab and CS MOS.

Instructions and scoring scale layout. Changes in the scoring interface layout may have influenced the scoring more than expected. The use of purely categorical scales in CS, displayed vertically and without a graphical indication of categories being equally spaced across the scale (as done in the lab interface, see Fig. 1), may have altered participants' scoring criterion. In addition, instructions (kept to a minimum in CS) may have not been clear enough. Thus, we hypothesized that *by repeating exactly the same experiment, with the same amount of instruction and the same scoring interface*, we would obtain similar aesthetic appeal MOS in Lab and CS.

Non-repeatability of aesthetic appeal assessments. Finally, we could not exclude the hypothesis that lab MOS, rather than CS ones, were unreliable. Despite previous evidence [20], aesthetic appeal assessments may not be repeatable at all. We hypothesized therefore that, *by running a new lab experiment*, we would find aesthetic appeal MOS in disagreement with both our previous Lab and CS results.

The first two hypotheses are closely related to the design of the CS task, whereas the latter two concern the design (and feasibility) of both the lab and the CS experiment. We investigate them in the following two sections, respectively.

4. ISSUES IN THE SETUP OF CS EXPERIMENTS

4.1 Bias due to the scoring interface

The co-occurring assessment of aesthetic appeal and recognizability could have affected the aesthetic appeal MOS, due to priming and/or worker sloppiness. To verify if that was the case, we repeated the original experiment in CS, this time presenting only one scale at the bottom of the image, either for rating aesthetic appeal or recognizability. We refer to this second CS experiment as the *Phase II* experiment, as opposed to the original one (referred to as Phase I).

Except for the change in task (single attribute scoring), the interface, experimental protocol and payment scheme were kept identical to Phase I (see Sec. 3). 26 campaigns (13 per attribute/task) were run and completed within 4 days. The scores provided by about 10% of the workers were excluded due to incorrect answering of content questions and unusual scoring time and behavior.

We evaluated whether the single attribute scoring had a positive effect on the repeatability of aesthetic appeal MOS by checking (1) whether the correlation between aesthetic appeal MOS obtained in lab and CS Phase II has increased, and (2) whether the (linear) relationship between aesthetic appeal and recognizability MOS was similar in CS Phase II and lab. Results are reported in Table 1, which shows that for both recognizability and aesthetic appeal, the correlation between lab and CS MOS is *smaller* than was found for the original CS experiment. This is also reflected in the relationship between aesthetics and recognizability in lab and in CS Phase II (Fig. 3). Despite the fact that workers scored one single attribute at a time, thus theoretically avoiding the unconscious link to the other attribute, MOS of recognizability and aesthetics became even more correlated (CS Phase II) than when the attributes were scored simultaneously (CS Phase I).

Hence, the carelessness or sloppiness of workers participating in the tests was not the reason for high dependency between aesthetic appeal and recognizability in CS results: a simple copying of recognizability scores into the aesthetic appeal scale was not possible in this new setup. However, 41% of the workers in fact participated in both aesthetic appeal and recognizability scoring campaigns (no limitations were envisioned to prevent this to happen). Consequently, unconscious bias still possibly existed for linking the two attributes. We then eliminated from our analysis the scores from all workers who participated in both tasks, and checked whether that was the case. More than 60% of the scores were filtered, as workers participating in both tasks conducted multiple campaigns for each task.

Table 1: Pearson correlation coefficients for MOS of aesthetic appeal and recognizability in the different experimental conditions.[2]

	Aesthetic Appeal MOS Correlations			
	CS-PII-R	CS-PII	CS-PII-S	CS-PI
Lab-PI	0.4358	0.4801	0.4461	0.5179
CS-PII-R	-	0.9574	0.9196	0.8713
CS-PII	-	-	0.9537	0.9112
CS-PII-S	-	-	-	0.8654
	Recognizability MOS Correlations			
	CS-PII-R	CS-PII	CS-PII-S	CS-PI
Lab-PI	0.8132	0.8188	0.8123	0.8678
CS-PII-R	-	0.9396	0.8758	0.8952
CS-PII	-	-	0.9268	0.9483
CS-PII-S	-	-	-	0.9077

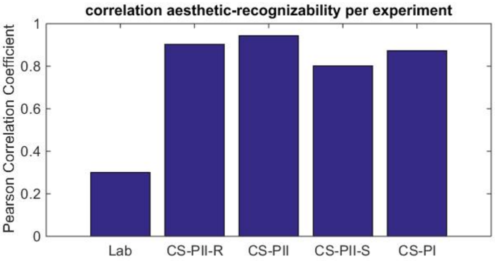


Figure 3: Pearson correlation coefficients between aesthetic appeal and recognizability in lab and in CS settings in Phase I and Phase II.[1]

By re-computing the correlations between CS Phase II and lab MOS (see Lab-PI vs. CS-PII-S in Table 1), we notice that for aesthetic appeal, correlation has dropped even further ($\rho = 0.45$) compared to the experiments with unfiltered duplicate workers ($\rho = 0.48$). On the other hand, the correlation between aesthetics and recognizability MOS decreased by about 17%, becoming slightly closer to the relationship observed in the lab. We conclude that there may have been a small unconscious linking bias due to the concurrent scoring of the two attributes together (simultaneous, as in the case of the CS Phase I experiment, or delayed in time, as in the case of CS Phase II); nevertheless, the lower correlation between recognizability and aesthetic appeal MOS

[1] MOS obtained in the lab are indicated as 'Lab', CS Phase I results as 'CS-PI', and Phase II results as 'CS-PII'. CS-PII MOS obtained excluding all workers who performed both tasks are indicated as CS-PII-S; Re-aligned MOS are indicated as CS-PII-R.

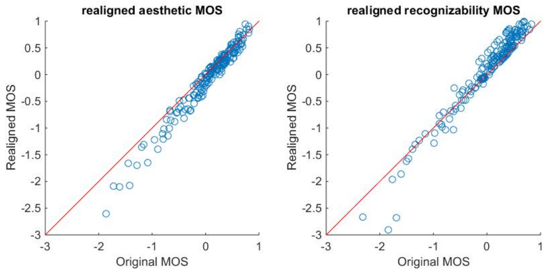


Figure 4: Realigned Aesthetic Appeal (Left) and Recognizability (Right) MOS.

may also depend on the imprecision of the MOS computed from a smaller amount of scores (7 on average).

4.2 Context effects and realignment

The absolute category rating (ACR) scale used in our experiments is known to present several drawbacks, among which a critical one is that of context effects [3, 4]. Context effects stem from the cognitive bias that leads subjects to use the entirety of a scoring scale (in case of ACR, until ‘bad’), to express the quality range that is visualized in the stimulus set. So, given a stimulus set having true quality values covering a range [0,A], and a second set of stimuli covering the range [A/2,A], it is quite likely that the worst stimulus of the second set will still obtain a MOS close to ‘bad’. Dividing the original set of 200 images into 13 campaigns possibly spanning different aesthetic appeal ranges, may have led to context effects, despite the addition of “anchor” images (see sec. 3).

To verify this, we re-aligned the MOS in all campaigns, as per [17]. We linearly transformed the MOS of the 5 anchor images scored in each campaign c, $MOS_c(A_i), i, = 1, ..., 5, c = 1, ..., 13$, into the scale of a reference campaign $c*$ and obtained the slope parameter α_C and the offset β_C per each campaign. As reference campaign $c*$, we chose the campaign spanning the widest range of MOS values. We then applied for each image I in campaign C the following linear transformation to obtain the realigned value $MOS^R_C(I)$ as expressed in the scale of the reference campaign $c*$: $MOS^R_C(I) = \alpha_c MOS_C(I) + \beta_c$.

The realigned MOS, for both recognizability and aesthetics, are depicted against the original MOS in Fig. 4. Although some contextual effects at the bottom end of the scale can be observed (original MOS were slightly compressed in range with respect to the realigned ones), for the most part, the re-alignment did not significantly change the magnitude and/or ordering of the MOS (Pearson correlation between original and realigned MOS was above 0.96 for both recognizability and aesthetic appeal). Re-aligning MOS did not significantly affect the relationship between lab and CS MOS (see Lab-PI vs. CS-PII-R in Table 1) or the relationship between aesthetic appeal and recognizability MOS, which are still highly correlated ($\rho > 0.9$) as shown in Fig. 3.

5. LAB-TO-CS REPEATABILITY ISSUES

In section 4 we showed that the lack of repeatability of our lab results in CS was not due to specific characteristics of the CS task. The problem may instead derive from either an intrinsic impossibility of repeating aesthetic appeal measurements, or a too high discrepancy between the lab and CS scoring tasks. To verify this, we ran a new set of experiment, both in controlled lab environment and CS, hereafter referred to as *Phase III*. The setup of this comparative experiment was nearly identical to that of Phase I (in terms of experimental protocol, methodology, worker reliability control mechanisms in CS, and environmental settings for the lab tests), except for the following changes:

Table 2: Correlation between MOS of aesthetic appeal collected in Phase I and III (lab and crowdsourcing).

Correlation	Lab-PI CS-PI	Lab-PI Lab-PIII	Lab-PIII CS-PIII	Lab-PI CS-PIII
Pearson	0.679	0.853	0.872	0.837

1. As we were concerned with the repeatability of aesthetic appeal measurements, we limited our investigation to this attribute only. We asked to rate the level of “beauty” of the image instead of the previously used “aesthetic appeal”. We expected this not to affect repeatability in a positive way per se, as we proved in Sec. 4 that scoring aesthetic appeal alone in CS did not increase the correlation between Lab and CS MOS.
2. We used exactly the same user interface for both lab- and CS-based assessments, including the same instructions and training process.
3. The scoring scale was presented in the same screen as the image (at the bottom). 5-point ACR scale was used again, displayed horizontally similarly to what was used in the previous lab experiment (see Fig. 1), but this time with the category labels attached to the scale ticks.
4. For efficiency purposes, a subset of 40 images from those used in previous phases that spanned the entire range of aesthetic appeal (according to the lab Phase I results), were selected for this new experiment. We added 14 new images for additional purposes detailed in [21], making the total number of images 54.

24 naïve subjects rated all 54 images in the lab, with settings identical to those of Phase I (see Sec. 3). For the CS experiment, images were split into three campaigns of 18 images each plus five anchor images to minimize contextual effects (see Sec.3). In each campaign, 30 workers were paid 0.50$, scoring 22 images in total. We screened lab participants for outliers according to [11] and CS workers for reliability following the same methodology as in the previous two phases [18]. One lab participant and about 25 % of CS workers were excluded from the results.

Results are reported in Table 2. A high correlation (0.872) is achieved between Lab and CS MOS of Phase III, with an increase of 22% with respect to Phase I. The lab Phase III MOS are highly correlated to the MOS collected in the original lab experiment, indicating that image aesthtetic appeal measurements are repeatable, and this repeatability is independent on the wording of the attribute to be scored (“aesthetic appeal” in Phase I and “level of beauty” in Phase III). This change in wording may have instead affected results in CS, as demonstrated by high correlations between CS Phase III and the original Lab Phase I results.

6. LESSONS LEARNT & CONCLUSIONS

We examined the issues related to repeating lab-based aesthetic appeal assessments in crowdsourcing. In an initial experiment, we replicated an existing lab experiment into a crowdsourcing setting, and found that the aesthetic appeal MOS obtained in CS were poorly correlated with those had in the lab. Our subsequent analysis showed that this outcome was *not due to sloppy task performance of the CS workers.* We hypothesized that, since in our original CS interface, recognizability and aesthetic appeal were to be scored in the same screen (and in this order), workers could copy-paste the recognizability score on the aesthetic appeal scale. By asking workers to score only one attribute at the time, we showed that the similarity between lab and CS aesthetic appeal MOS remained low, and the correlation between recognizability and aesthetics in CS stayed high. We were also able to verify that the segmentation of the image set in 13 CS campaigns did not have a negative influence on repeatability. Re-aligned CS aesthetic appeal MOS were just as correlated with the lab ones as the non-realigned ones. This result indicates that context effects were negligible in our experiments, despite the large number of image subsets involved in the assessments. This suggests that *the use of anchor images is beneficial in keeping context effects to a minimum*, allowing to merge MOS from different campaigns on a single scale for a minimum overhead.

In a final experiment, we ran a new lab- and CS-based assessment of a subset of the previous images. Here we kept the scoring interfaces identical between lab and CS. The results showed that (1) the odd results obtained in the original experiment *were not due to the impossibility of repeating aesthetic appeal measurements*, as we showed that MOS of aesthetic appeal obtained in the two distinct lab experiments were highly correlated; and (2) the presentation of the scoring scale, purely categorical in CS and with a hint of linearity in the Lab, played a major role in the discrepancies found for the first experiment. This may have led workers to position images in different portions of the scale. In addition, in our experiments, *the use of ACR labels along a linear scale* (thus, with equally spaced tick marks), would provide the most repeatable results. A possible reason for this could reside in the fact that purely categorical scales are subject to individual participant's interpretation of where the boundary across the category lies [6], whereas this effect is minimized by graphically indicating equal width of categories as done in this last experiment.

7. REFERENCES

[1] K.-T. Chen, C.-J. Chang, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. Quadrant of euphoria: a crowdsourcing platform for QoE assessment. *Network, IEEE*, 24(2):28–35, Mar. 2010.

[2] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. In *ACM international conference on Multimedia*, MM '09, pages 491–500. ACM, 2009.

[3] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach. All subjective scales are not created equal: The effects of context on different scales. *Signal processing*, 77(1):1–9, 1999.

[4] H. de Ridder. Cognitive issues in image quality measurement. *Journal of Electronic Imaging*, 10(1):47–55, 2001.

[5] A. Doan, R. Ramakrishnan, and A. Y. Halevy. Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4):86–96, 2011.

[6] P. G. Engeldrum. *Psychometric scaling: a toolkit for imaging systems development.* Imcotek Press, 2000.

[7] T. Hossfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel. Best practices and recommendations for crowdsourced qoe v1.0. *COST Action IC1003 Qualinet*, Oct. 2014.

[8] T. Hoßfeld and C. Keimel. Crowdsourcing in QoE Evaluation. In S. Möller and A. Raake, editors, *Quality of Experience: Advanced Concepts, Applications and Methods.* Springer, 2014.

[9] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia. Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing. *IEEE Transactions on Multimedia*, 16, 2014.

[10] P. Isola, J. Xiao, A. Torralba, and A. Oliva. What makes an image memorable? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 145–152. IEEE, 2011.

[11] ITU-R BT.500. Methodology for the subjective assessment of the quality of television pictures, 2002.

[12] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.

[13] C. Keimel, J. Habigt, C. Horch, and K. Diepold. Qualitycrowd - a framework for crowd-based quality evaluation. In *Picture Coding Symposium (PCS), 2012*, pages 245–248, May 2012.

[14] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

[15] P. Le Callet, S. Möller, and A. Perkis. Qualinet white paper on definitions of quality of experience (2012).

[16] W. A. Mansilla, A. Perkis, and T. Ebrahimi. Implicit experiences as a determinant of perceptual quality and aesthetic appreciation. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 153–162. ACM, 2011.

[17] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, and P. Le Callet. Aligning subjective tests using a low cost common set. In *Euro ITV*, 2011.

[18] J. Redi, T. Hossfeld, P. Korshunov, F. Mazza, I. Povoa, and C. Keimel. Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal. In *ACM CrowdMM 2013*, Barcelona, Spain, Oct. 2013.

[19] J. Redi, I. Povoa, et al. The role of visual attention in the aesthetic appeal of consumer images: A preliminary study. In *Visual Communications and Image Processing (VCIP), 2013*, pages 1–6. IEEE, 2013.

[20] J. A. Redi, Y. Zhu, H. de Ridder, and I. Heynderickx. How passive image viewers became active multimedia users. In *Visual Signal Quality Assessment*, pages 31–72. Springer, 2015.

[21] E. Siahaan, J. Redi, and A. Hanjalic. Beauty is in the scale of the beholder: Comparison of methodologies for the subjective assessment of image aesthetic appeal. In *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pages 245–250, Sept 2014.